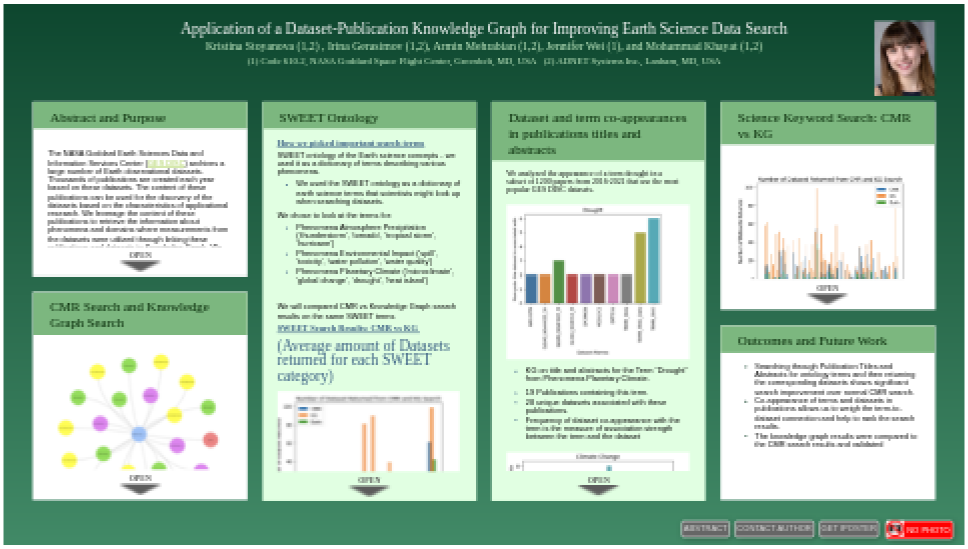
# Application of a Dataset-Publication Knowledge Graph for Improving Earth Science Data Search


Application of a Dataset-Publication Knowledge Graph for Improving Earth Science Data Search

Kristina Stoyanova (1,2), Irina Gerasimov (1,2), Armin Mehrabian (1,2), Jennifer Wei (1), and Mohammad Khayat (1,2)

Abstract and Purpose

CMR Search and Knowledge Graph Search

SWEET Ontology

(Average amount of Datasets returned for each SWEET category)

Dataset and term co-appearances in publications titles and abstracts

Science Keyword Search: CMR vs KG

Outcomes and Future Work

Kristina Stoyanova (1,2) , Irina Gerasimov (1,2), Armin Mehrabian (1,2), Jennifer Wei (1), and Mohammad Khayat (1,2)

(1) Code 610.2, NASA Goddard Space Flight Center, Greenbelt, MD, USA (2) ADNET Systems Inc., Lanham, MD, USA

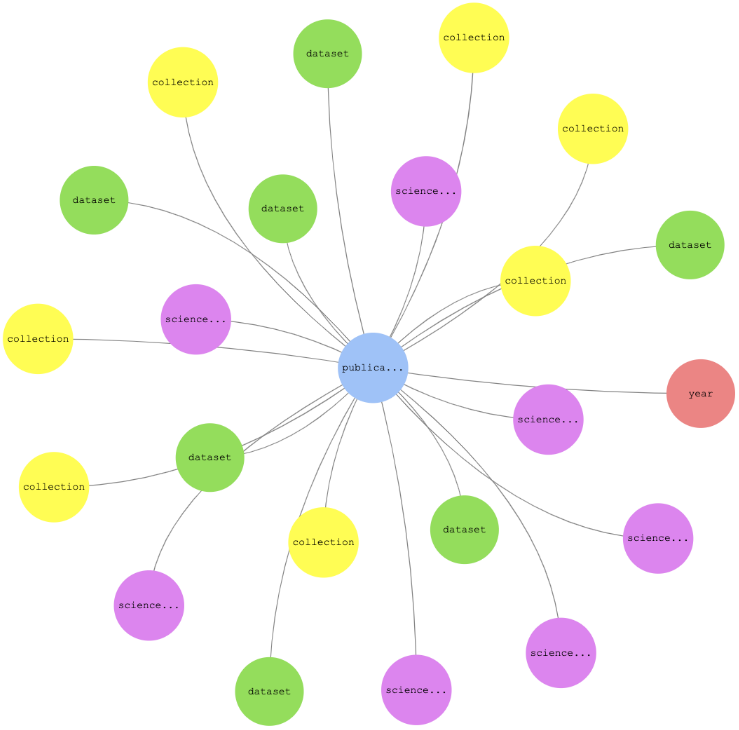
PRESENTED AT:

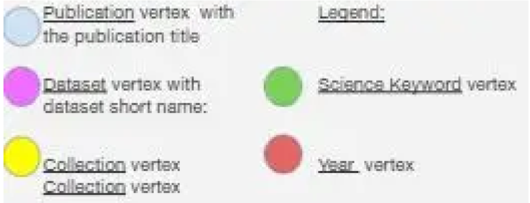
# ABSTRACT AND PURPOSE

The NASA Goddard Earth Sciences Data and Information Services Center (GES DISC (http://disc.gsfc.nasa.gov/)) archives a large number of Earth observational datasets. Thousands of publications are created each year based on these datasets. The content of these publications can be used for the discovery of the datasets based on the characteristics of applicational research. We leverage the content of these publications to retrieve the information about phenomena and domains where measurements from the datasets were utilized through linking these publications and datasets in Knowledge Graph. We retrieve phenomena and domain information using SWEET (http://github.com/ESIPFed/sweet) (Semantic Web for Earth and Environmental Terminology) ontology and produce the set of keywords that are linked to the datasets. Further, we evaluate this link strength according to the frequency of dataset usage in the papers mentioning these keywords. We demonstrate how this linkage can improve dataset search by comparing the search results obtained from the Common Metadata Repository (CMR (https://cmr.earthdata.nasa.gov/search)) search and publications-based data.

See more about the KG here (http://agu2021fallmeeting-agu.ipostersessions.com/Default.aspx?s=2E-97-8B-0F-BA-71-A4-CA-66-F0-5C-22-F1-6D-ED-1C)

# CMR SEARCH AND KNOWLEDGE GRAPH SEARCH




🔵 Publication vertex with the publication title

Legend:

🟣 Dataset vertex with dataset short name:

🟢 Science Keyword vertex

🟡 Collection vertex
Collection vertex

🔴 Year vertex

Create relevant vertices

- Ex: Publications, Datasets, Science Keywords

Edges connect vertices

- Ex: CreatedBy Edges

Our KG abstract and title search provide an insight into how the full knowledge graph can help us to improve the search.

- The publication vertex may have an attribute of a title or abstract that contains an ontology term, which can then connect that ontology term to a dataset. Which is what our KG search is doing.

# SWEET ONTOLOGY

**How we picked important search terms**

SWEET ontology of the Earth science concepts - we used it as a dictionary of terms describing various phenomena.

- We used the SWEET ontology as a dictionary of earth science terms that scientists might look up when searching datasets.

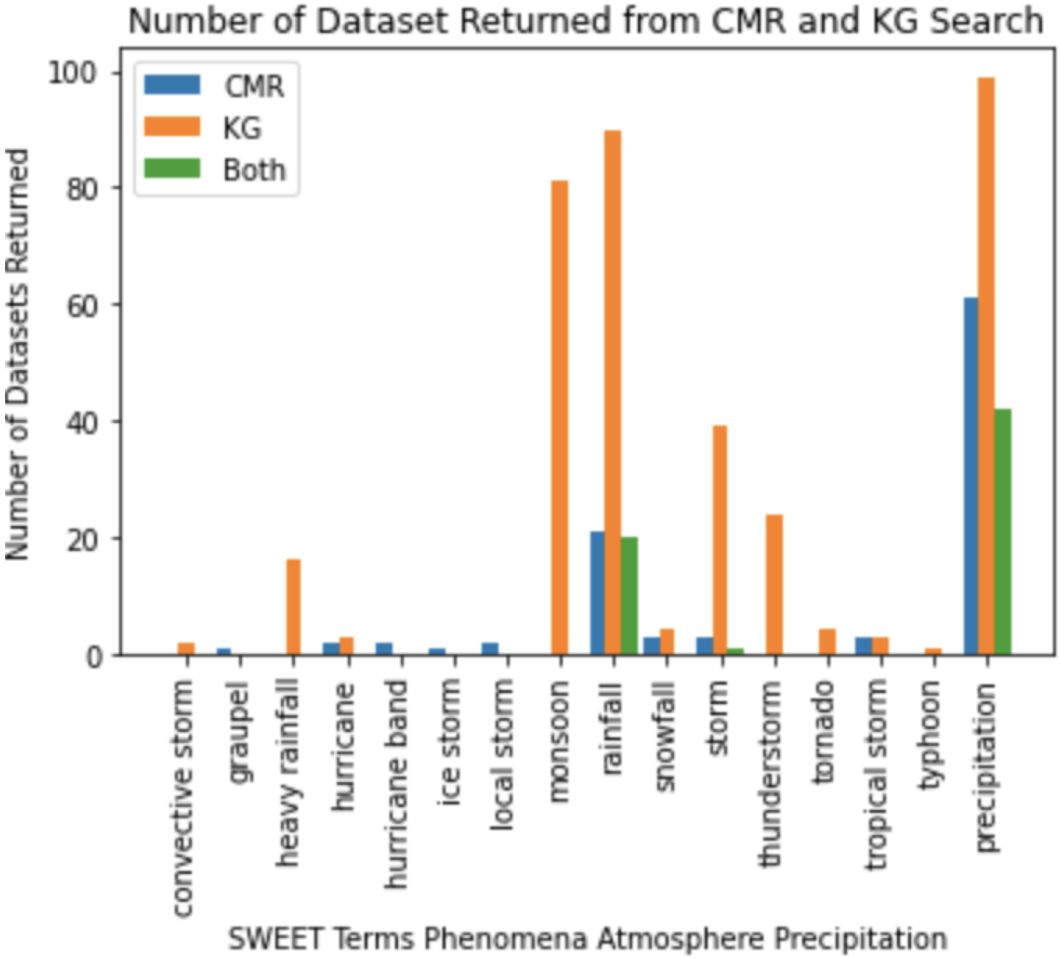
We chose to look at the terms for:

- Phenomena Atmosphere Precipitation ('thunderstorm', 'tornado', 'tropical storm', 'hurricane')
- Phenomena Environmental Impact ('spill', 'toxicity', 'water pollution', 'water quality')
- Phenomena Planetary Climate ('microclimate', 'global change', 'drought', 'heat island')

We will compared CMR vs Knowledge Graph search results on the same SWEET terms

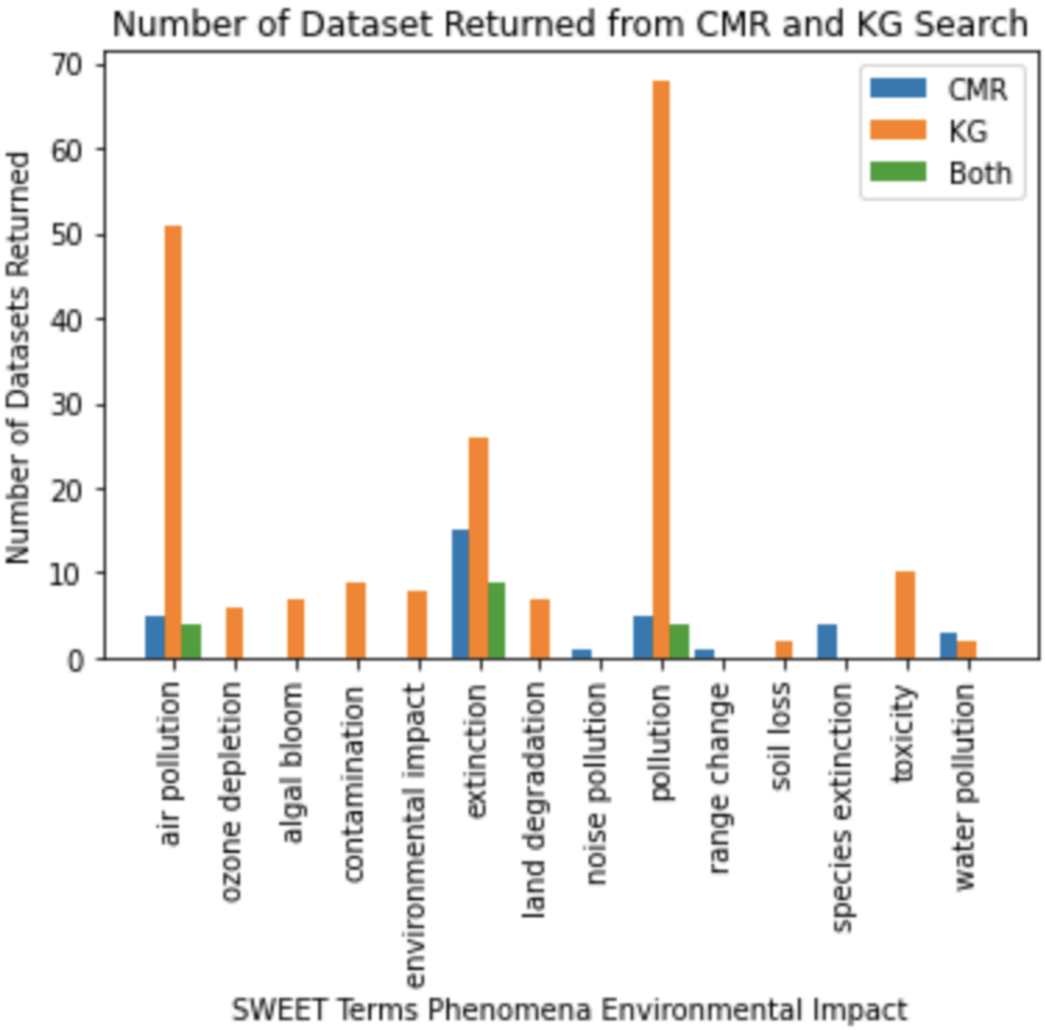
**SWEET Search Results: CMR vs KG**

(Average amount of Datasets returned for each SWEET category)



- We compared CMR and the KG on 48 SWEET terms
- Overall the KG returned more datasets than CMR

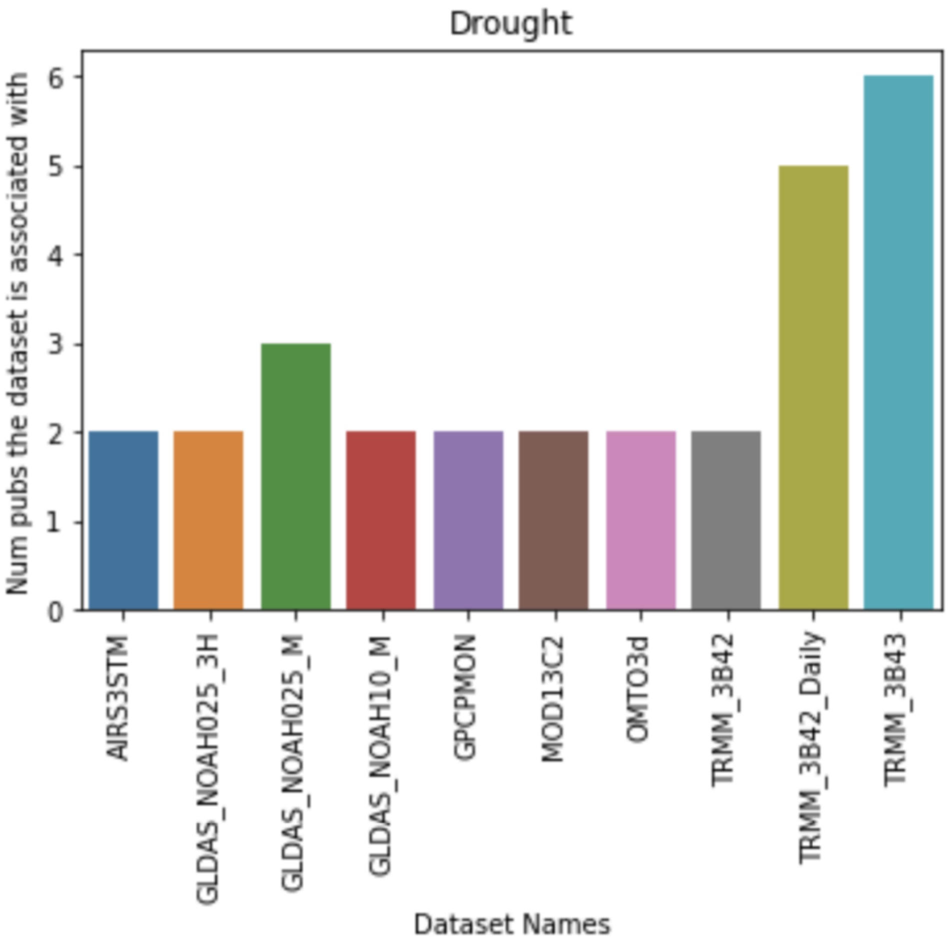
- For most of the terms, the KG returns a set of datasets that include all the ones from CMR



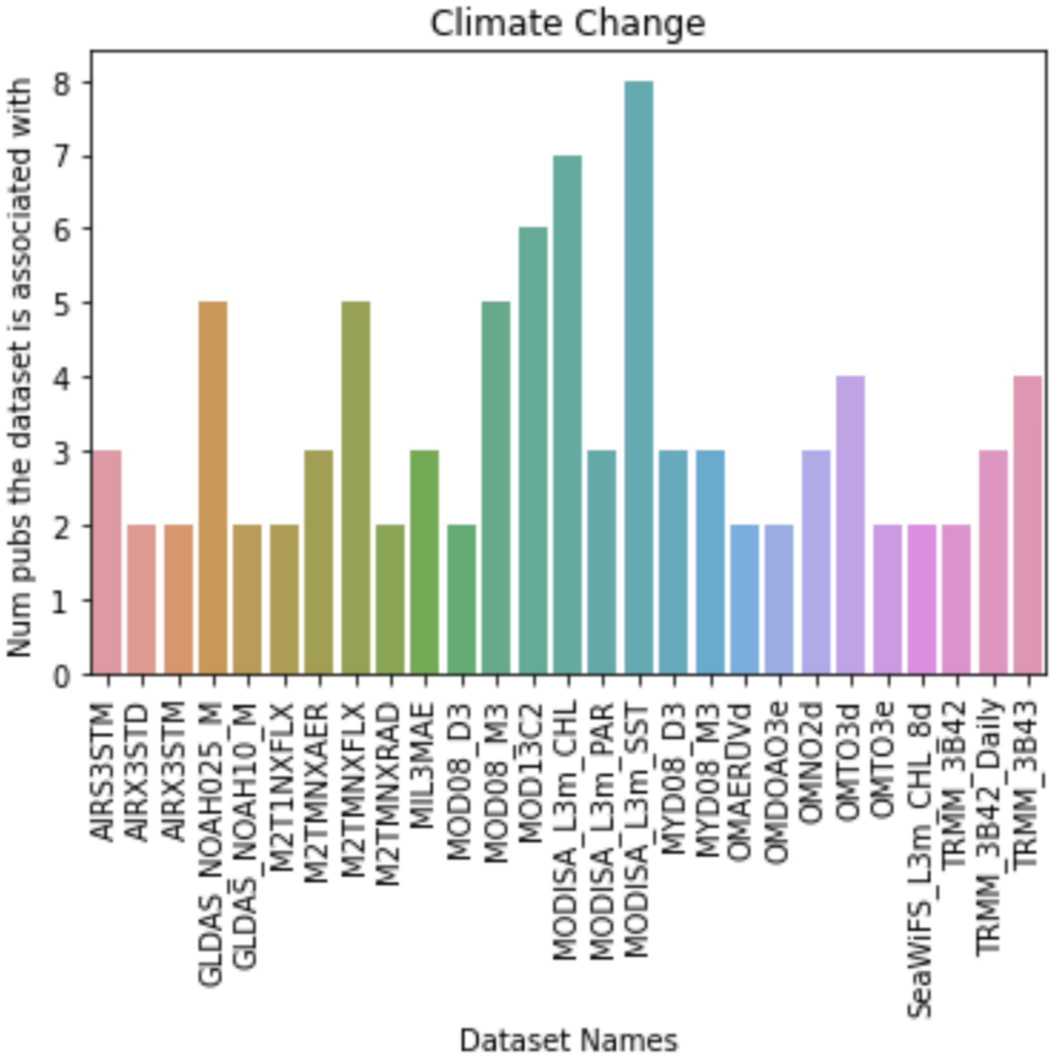
- Enhancing CMR search with results from the KG can capture more term dataset relationships and return results for words that were previously not queriable on CMR

# DATASET AND TERM CO-APPEARANCES IN PUBLICATIONS TITLES AND ABSTRACTS

We analyzed the appearance of a term drought in a subset of 1200 papers from 2016-2021 that use the most popular GES DISC datasets.



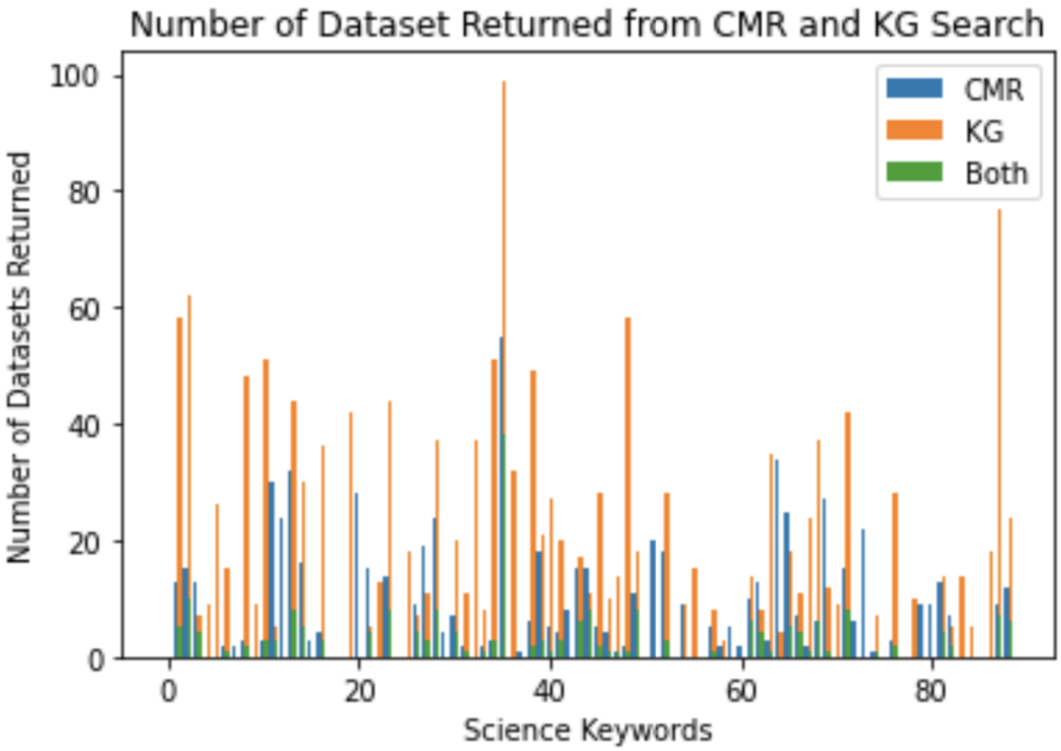
- KG on title and abstracts for the Term “Drought” from Phenomena Planetary Climate.
- 19 Publications containing this term.
- 28 unique datasets associated with these publications.
- Frequency of dataset co-appearance with the term is the measure of association strength between the term and the dataset



- Term: “Climate Change” from Phenomena Planetary Climate
- 50 Publications
- 65 unique datasets associated with these publications
- From 2016 - 2021
- Enabling usage-based discovery: search for datasets in paper titles and abstracts by data usage terms.

KG search returns not only the number of unique datasets in publications that have the term in their title or abstract but also the number of times each dataset is used in multiple publications. The count of number times a dataset is used in a publication can be used as weights in the graph for usage-based dataset discovery.

# SCIENCE KEYWORD SEARCH: CMR VS KG



- Average KG:  17.8 datasets
- Average CMR: 19.7 datasets
- On Average, the KG returned 90% of the datasets that CMR returned.

- We validated the KG results for the GCMD science keywords. The KG returns 90% of the datasets that CMR returns. At this moment the KG only had 1200 publications, we expect the results to increase once more publications are added to the KG.
- We also compared the KG search and CMR search on 90 Scientific Keywords from the KG, which are words scientists have created to describe CMR datasets.
- CMR search uses the GCMD science keywords, and interesting the KG was inclusive of the results and returned more.

# OUTCOMES AND FUTURE WORK

- Searching through Publication Titles and Abstracts for ontology terms and then returning the corresponding datasets shows significant search improvement over normal CMR search.
- Co-appearance of terms and datasets in publications allows us to weigh the term-to-dataset connection and help to rank the search results.
- The knowledge graph results were compared to the CMR search results and validated

# ABSTRACT

Finding a dataset at a NASA data center that is the best fit for the researcher’s application presents a challenge, not only for a novice user but for an experienced one, due to the data complexity and a multitude of choices of the existing data. Users often search for the data based on the application they are interested in, their research domain, phenomena, research topic, etc. As existing dataset metadata may not cover these search terms, the user may not obtain the most relevant results for their purpose. This problem was addressed by leveraging the content of the titles and abstracts of the research papers that utilize NASA datasets. For this, features from the paper titles and abstracts were extracted, and then a knowledge graph (KG) was used to link these features to the datasets used in that paper. The search for the datasets was tested by querying this knowledge graph through various terms extracted from Earth Science ontologies such as Semantic Web for Earth and Environment Technology (SWEET), and it was shown that this KG search outperforms the existing search that exclusively queries the dataset metadata.